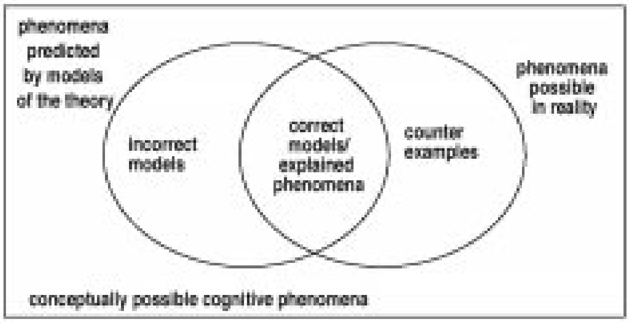
phenomena predicted by models of the theory
phenomena possible in reality
incorrect models
correct models/ explained phenomena
counter examples
conceptually possible cognitive phenomena

Figure 1 (Taatgen). Diagram to illustrate successes and problems of a theory of cognition.

Figure 1 illustrates the issue. Consider the set of all possibly conceivable cognitive phenomena, of which only a subset contains phenomena that can actually occur in reality. Then the goal of a theory is to predict which of the conceivable phenomena are actually possible, and the success of a theory depends on the overlap between prediction and reality. The problems of a theory can be found in two categories: counterexamples, phenomena that are possible in reality but are not predicted by the theory, and incorrect models, predictions of the theory that are not possible in reality. The issue of incorrect models is especially important, because an unrestricted Turing Machine is potentially capable of predicting any conceivable cognitive phenomenon. One way to make the Newell Test more precise would be to stress the falsifiability aspects for each of the items on the test. For some items this is already more or less true in the way they are formulated by Anderson & Lebiere (A&L), but others can be strengthened, for example:

Flexible behavior. Humans are capable of performing some complex tasks after limited instructions, but other tasks first require a period of training. The theory should be able to make this distinction as well and predict whether humans can perform the task right away or not.

Real-time performance. The theory should be able to predict human real-time performance, but should not be able to predict anything else. Many theories have parameters that allow scaling the time predictions. The more these parameters are present, the weaker is the theory. Also the knowledge (or network layout) that produces the behavior can be manipulated to adjust time predictions. Restricting the options for manipulation strengthens the theory.

Knowledge integration. One property of what A&L call "intellectual combination" is that there are huge individual differences. This gives rise to the question how the theory should cope with individual differences: Are there certain parameters that can be set that correspond to certain individual differences (e.g., Lovett et al. 1997; Taatgen 2002), or is it mainly a difference in the knowledge people have? Probably both aspects play a role, but it is of chief importance that the theory should both predict the breadth and depth of human behavior (and not more).

Use natural language. The theory should be able to use natural language but should also be able to assert what things cannot be found in a natural language. For example, the ACT-R model of learning the past tense shows that ACT-R would not allow an inflectional system in which high-frequency words are regular and low-frequency words are irregular.

Learning. For any item of knowledge needed to perform some behavior, the theory should be able to specify how that item has been learned, either as part of learning within the task, or by showing why it can be considered as knowledge that everyone has. By demanding this constraint on models within a theory, models that have unlearnable knowledge can be rejected. Also, the learning system should not be able to learn knowledge that people cannot learn.

Development. For any item of knowledge that is not specific to a certain task, the theory should be able to specify how that item of knowledge has been learned, or to supply evidence that that item of knowledge is innate. This constraint is a more general version of the learning constraint. It applies to general strategies like problem solving by analogy, perceptual strategies, memorization strategies, and the like.

Another aspect that is of importance for a good theory of cognition is parsimony. This is not an item on Newell's list, because it is not directly tied to the issue of cognition, but it was an important aspect of Newell's research agenda. This criterion means that we need the right number of memory systems, representations, processing, and learning mechanisms in the theory, but not more. An advantage of parsimony is that is makes a stronger theory. For example, SOAR has only one learning mechanism, chunking. This means that all human learning that you want to explain with SOAR has to be achieved through chunking, as opposed to ACT-R, which has several learning mechanisms. Of course, SOAR's single mechanism may eventually be found lacking if it cannot account for all human learning.

To conclude, research in cognitive modeling has always had a positivistic flavor, mainly because it is already very hard to come up with working models of human intelligence in the first place. But as cognitive theories gain in power, we also have to face the other side of the coin: to make sure that our theories rule out wrong models. This is not only an issue for philosophers of science, but a major issue if we want to apply our theories in human-computer interaction and education. There, it is of vital importance that we should be able to construct models that can provide reliable predictions of behavior without having to test them first.

Cognitive architectures have limited explanatory power

Prasad Tadepalli

School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331-3202. **tadepall@cs.orst.edu**
http://www.eecs.orst.edu/~tadepall

Abstract: Cognitive architectures, like programming languages, make commitments only at the implementation level and have limited explanatory power. Their universality implies that it is hard, if not impossible, to justify them in detail from finite quantities of data. It is more fruitful to focus on particular tasks such as language understanding and propose testable theories at the computational and algorithmic levels.

Anderson & Lebiere (A&L) undertake the daunting task of evaluating cognitive architectures with the goal of identifying their strengths and weaknesses. The authors are right about the risks of proposing a psychological theory based on a single evaluation criterion. What if the several micro-theories proposed to meet different criteria do not fit together in a coherent fashion? What if a theory proposed for language understanding and inference is not consistent with the theory for language learning or development? What if a theory for playing chess does not respect the known computational limits of the brain? The answer, according to Newell, and A&L, is to evaluate a cognitive theory along multiple criteria such as flexibility of behavior, learning, evolution, knowledge integration, brain realization, and so forth. By bringing in multiple sources of evidence in evaluating a single theory, one is protected from *overfitting,* a problem that occurs when the theory has too many degrees of freedom relative to the available data. Although it is noncontroversial when applied to testable hypotheses, I believe that this research strategy does not work quite as well in evaluating cognitive architectures.

Science progresses by proposing testable theories and testing them. The problem with cognitive architectures is that they are not theories themselves but high-level languages used to imple-

https://doi.org/10.1017/S0140525X03400137 Published online by Cambridge University Press

ment theories, with only some weak architectural constraints. Moreover, these languages are computationally universal and thus are equivalent to one another in the sense that one language can simulate the other. How does one evaluate or falsify such universal languages? Are the multiple criteria listed by the authors sufficient to rule out anything at all, or do they simply suggest areas to improve on? The authors' grading scheme is telling in this respect. It only evaluates how an architecture satisfies one criterion better than another criterion, and does not say how to choose between two architectures. One cannot, of course, duck the question merely by choosing an architecture based on the criterion one is interested in explaining. This is precisely the original problem that Newell was trying to address through his multiple criteria.

The authors suggest that timing constraints and memory limitations imply that one cannot only program arbitrary models in ACT-R. But that still leaves room for an infinite variety of models, and ACT-R cannot tell us how to choose between them. To take an analogy to programming languages: It is possible to design an infinite variety of cognitive architectures and implement an infinite variety of models in each one. Can we ever collect enough evidence to be able to choose one over another?

This suggests to me that a cognitive theory must be carefully distinguished from the concrete implementation and the underlying architecture. Just as a programming language can implement any given algorithm, a cognitive architecture can instantiate any cognitive theory (albeit with some variations in time efficiencies). This should not count as evidence for the validity of the architecture itself, any more than good performance of an algorithm should count as evidence for the validity of the programming language. Cognitive science can make better progress by carefully distinguishing the algorithm from the architecture and confining the claims to those parts of the algorithm that are in fact responsible for the results. Consider, for example, ACT-R's theory of past-tense learning by children. More specifically, consider the empirical observation that the exceptions tend to be high-frequency words. A&L attribute this to the fact that only high-frequency words develop enough base-level activation to be retrieved in ACT-R. In more general terms, only high-frequency words provide sufficient training data for the system to be able to learn an exception. How much of this explanation is a result of the particulars of ACT-R theory as opposed to being a necessary consequence of learning in a noisy domain? If any learning system that operates in a noisy environment needs more training data to learn an exception, why should this be counted as evidence for the ACT-R theory? Similar criticisms can be leveled against other cognitive architectures and mechanisms such as SOAR and chunking, connectionism, and backprop.

In other words, even when multiple criteria are used to evaluate a cognitive architecture, there still remains an explanatory gap (or a leap of faith) between the evidence presented and the paradigm used to explain it. To guard against such over-interpretation of the evidence, Ohlsson and Jewett propose "abstract computational models," which are computational models that are designed to test a particular hypothesis without taking a stand on all the details of a cognitive architecture (Ohlsson & Jewett 1997). Similar concerns are expressed by Pat Langley, who argues that the source of explanatory power often lies not in the particular cognitive architecture being advanced but in some other fact such as the choice of features or the problem formulation (Langley 1999). Putting it another way, there are multiple levels of explanations for a phenomenon such as past-tense learning or categorization, including computational theory level, algorithmic level, and implementation level. Computational theory level is concerned with *what* is to be computed, whereas algorithmic level is concerned with *how* (Marr 1982). Cognitive architecture belongs to the implementation level, which is below the algorithmic level. Where the explanatory power of an implementation mostly lies is an open question.

Only by paying careful attention to the different levels of explanations and evaluating them appropriately can we discern the truth. One place to begin is to propose specific hypotheses about the algorithmic structure of the task at hand and evaluate them using a variety of sources of evidence. This may, however, mean that we have to put aside the problem of evaluating cognitive architectures, for now or forever.

ACKNOWLEDGMENTS
I thank Sridhar Mahadevan and Pat Langley for influencing my thinking on this matter and for their comments on an earlier draft.

Cognitive modelling of human temporal reasoning

Alice G. B. ter Meulen
Center for Language and Cognition, University of Groningen, 9700 AS Groningen, The Netherlands. **atm@let.rug.nl** **http://atm.nemit.net**

Abstract: Modelling human reasoning characterizes the fundamental human cognitive capacity to describe our past experience and use it to form expectations as well as plan and direct our future actions. Natural language semantics analyzes dynamic forms of reasoning in which the real-time order determines the temporal relations between the described events, when reported with telic simple past-tense clauses. It provides models of human reasoning that could supplement ACT-R models.

Real-time performance, the second criterion for a human cognitive architecture in Newell (1990), requires the system to operate as fast (or as slow) as humans (target article, sect. 2, Table 1) on any cognitive task. Real time is hence considered a constraint on learning as well as on performance (sect. 5). Although I certainly consider it an advantage of the ACT-R system that it does not rely on artificial assumptions about presentation frequency in the way classical connectionist systems do (Taatgen & Anderson 2002), the limited focus the two systems share on the acquisition of the morphological variability in the simple past-tense inflection in English ignores its obvious common semantic properties, which also must be learned. In this commentary, I propose to include in real-time performance the characteristic human ability to use time effectively when using language to encode information that systematically depends on contextual parameters, such as order of presentation or time of utterance.

Human linguistic competence includes automated processes of temporal reasoning and understanding, evidence of which is presented in our linguistic intuitions regarding the temporal relations that obtain between events described in coherent discourse. The presentation order in which simple past-tense clauses are produced in real time often contains important clues for the correct interpretation. As opposed to the past progressive (*John was leaving*) and the past perfect (*John had left*), the English simple past tense (*John left*) refers to an event that not only precedes the time of utterance but also is temporally located with respect to other events described by prior discourse. The following examples, (1) and (2), show that the order of presentation affects our understanding of what happened.

(1) *John lit a cigarette. He left.*
(2) *John left. He lit a cigarette.*

From (1) we understand that John left after he had lit a cigarette. But (2) makes us understand that the described events occurred in the opposite order. Obviously, the real-time order of presentation in this case determines the temporal relations between the events described. But this is not always so, as we see from examples (3) and (4), where reversing the order of the simple past-tense clauses does not affect the temporal relations between the events.

(3) *John slept for hours. He dreamt of Mary.*
(4) *John dreamt of Mary. He slept for hours.*

Either (3) or (4) makes us understand that John dreamt of Mary while he slept, which is reinforced by the lexical presupposition of dreaming requiring that the dreamer be asleep.

https://doi.org/10.1017/S0140525X03400137 Published online by Cambridge University Press